REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
06-02-2015	Conference Proceeding	-

4. TITLE AND SUBTITLE
Empirical Evaluation of Different Feature Representations for Social Circles Detection

5a. CONTRACT NUMBER
W911NF-14-1-0254

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Jesús Alonso, Roberto Paredes, Paolo Rosso

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Universitat Politècnica De València
Technology Transfer Office_CTT
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
-

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65349-MA.2

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Social circles detection is a special case of community detection in social network that is currently attracting a growing interest in the research community. We propose in this paper an empirical evaluation of the multi-assignment clustering method using different feature representation models. We define different vectorial representations from both structural egonet information and user profile features. We study and compare the performance on the available labelled Facebook data from the Kaggle competition on learning social circles in networks. We compare our results with several different baselines

15. SUBJECT TERMS
social circles detection, community detection, feature representations

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Paolo Rosso
UU	UU	UU	UU		19b. TELEPHONE NUMBER 963-877-007e

Report Title

Empirical Evaluation of Different Feature Representations for Social Circles Detection

ABSTRACT

Social circles detection is a special case of community detection in social network that is currently attracting a growing interest in the research community. We propose in this paper an empirical evaluation of the multi-assignment clustering method using different feature representation models. We define different vectorial representations from both structural egonet information and user profile features. We study and compare the performance on the available labelled Facebook data from the Kaggle competition on learning social circles in networks. We compare our results with several different baselines.

Conference Name: 7th Iberian Conference on Pattern Recognition and Image Analysis

Conference Date: June 16, 2015

Empirical Evaluation of Different Feature Representations for Social Circles Detection

Jesús Alonso, Roberto Paredes, and Paolo Rosso

Pattern Recognition and Human Language Technologies Research Center
Universitat Politècnica de València
https://www.prhlt.upv.es/

Abstract. Social circles detection is a special case of community detection in social network that is currently attracting a growing interest in the research community. We propose in this paper an empirical evaluation of the multi-assignment clustering method using different feature representation models. We define different vectorial representations from both structural egonet information and user profile features. We study and compare the performance on the available labelled Facebook data from the Kaggle competition on learning social circles in networks. We compare our results with several different baselines.

Keywords: social circles detection, community detection, feature representations

1 Introduction

Nowadays, users in social networks tend to organize the contacts in their personal networks by means of social circles, a tool already implemented by the major companies, like for instance Facebook lists or Google+ circles. However, this labelling is still mostly done manually and therefore a growing interest has risen in the automatic detection of these circles. In addition, this problem is related to the more general task of community detection in graphs, or the identification of subnetworks in a given network. The main difference between both problems is the use of information from users' profiles, apart from information from the network structure itself.

Despite the lack of a precise and well-accepted definition of community, there is a wide variety of methods and techniques designed to cope with community detection [3, 10]. Moreover, some techniques specifically designed for social circles detection are being developed currently [6, 7]. In this article, we present our approach based on multi-assignment clustering (MAC) [13, 4], originally a clustering technique for Boolean vectorial data not necessarily related to networks or graphs. The advantage of this technique is the possibility to assign the same object into several different clusters, different social circles. MAC has already been tried for social circles detection [6, 7] but only using a very simple feature representation, considering only user profile features, ignoring the network

structure. In our work we propose different and novel approaches by considering different representations of both network structure and user profile features.

The rest of the paper is structured as follows. In section two, we present previous works on community detection and social circles detection. In section three, we describe thoroughly our methodology, including the different data representations proposed and the baseline methods to compare with. In section four, we present the dataset and the evaluation measure of our experiments. In section five, we discuss the obtained results. Finally, we draw some conclusions.

2 Previous Work

2.1 Community Detection in Networks

From an abstract point of view, a network is equivalent to a graph, defined by a set of nodes connected by edges. Nevertheless, from the point of view of researchers devoted to a diversity of fields, the concept of network has additional connotations. Networks can represent real structures such as social networks, biological networks (neural synaptic networks, metabolical networks), technological networks (the Internet, the World Wide Web), logistic networks (distribution networks), etc. There is no well-accepted formal definition of community in general networks. However, there is a consensus on the fact that it consists of a group of nodes that are more densely connected to each other than to the nodes outside. The relation of membership in a community usually has an extra meaning, and the vertices in a community will probably share common properties or play similar roles within the graph.

Community detection is the task of automated identification of the communities of a network. A considerable number of methods have been developed to solve this problem [3, 10].

In real networks nodes are often shared among different communities. The most popular technique to detect overlapping communities is the clique percolation method [8]. Given a graph, a k-clique is defined as a complete subgraph of size k. Clique percolation consists in the identification of k-clique communities, defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. Despite of the good performance of this technique, clique percolation remains a hard computational problem, new and improved implementations still scale worse than some other overlapping community finding algorithms.

2.2 Application to Social Networks and Social Circles Detection

The study of social networks is a research topic with a history of decades and it has been recently revitalized by the appearance of new information and communication technologies which have opened new ways of interacting. Clustering of this social content has been studied designing several procedures. Some approaches base the clustering on the network links [10], while others consider

the semantic content of social interactions [15]. In between both methodologies, there has also been work on combining the links and the content for doing the clustering [9, 12]. Very recently, a new technique studied the characteristics of community structures formed around topical discussion clusters, using modularity maximization algorithms [2].

Social circles detection is a special case of this framework. Within a social network, an ego network or egonet is defined as the subgraph of the contacts of a particular user (called the ego). Thus, it includes all the contacts of the ego and the contact relationship between every pair of them. Then, the social circles of an ego can be considered as clusters of the egonet. Social circles may overlap (share nodes), for example university friends who were high school friends as well; and they may also present hierarchical inclusion (the nodes of a circle totally included into another), for example university friends into a generic friends category. Apart from the links of the egonet, user profile information is also normally considered in this task. The latest works on social circles detection define a generative model that considers circle memberships and a circle-specific profile similarity metric [6, 7].

3 Methodology

3.1 Multi-Assignment Clustering

Multi-Assignment Clustering (MAC) [13, 4] is a clustering method, originally developed for Boolean vectorial data, which allows for the possibility to assign the same object into several different clusters. It provides a decomposition of the data matrix $\mathbf{x}$ into a matrix containing the clusters prototypes $\mathbf{z}$ and a matrix representing the degree to which a particular data vector belongs to the different clusters $\mathbf{y}$. Finding optimal matrices $\mathbf{z}$ and $\mathbf{y}$ is NP-hard [14], but a probabilistic representation allows to drastically simplify the optimization problem. In [13] the authors propose to model the probability of x_{ij} under the signal model as:

$$p(x_{ij}|\mathbf{z},\beta) = \left[1 - \prod_{k=1}^{K} \beta_{kj}^{z_{ik}}\right]^{x_{ij}} \left[\prod_{k=1}^{K} \beta_{kj}^{z_{ik}}\right]^{1-x_{ij}}, \text{ where } \beta_{kj} := p(y_{kj} = 0) \quad (1)$$

In addition to the signal model, there is a noise model for the difference between the original data and the reconstruction made from $\mathbf{z}$ and $\mathbf{y}$. The model parameters are inferred by deterministic annealing [11, 1]. When MAC is applied to social circles detection, $\mathbf{y}$ is the matrix indicating which users belong to the different clusters, social circles.

In [6, 7] MAC was already employed and considered as a baseline method for social circles detection, although using only user profile information. This piece of evidence that MAC is a state-of-the-art technique, having recent and influential publications, helped us making the choice over alternative soft-clustering strategies. In this work, we propose to explore further its possibilities for this

task, investigating novel representations. We defend the fact that this technique still has potential and better results can be obtained. Furthemore, MAC is more adequate for large networks than other methods with a very high computational cost, like clique percolation.

As a novelty, we model the structural information of the egonets into diverse vectorial representations ready to be supplied to the algorithm. Several vectorial representations for user profile features were developed as well. Unlike the original MAC, we allow the input to be real data in $[0,1]^n$ as a way to model a hierarchy of link levels in the case of structural information, or an aggregation of the number of feature values shared by two users profiles in the case of user profile information.

In all the experiments, the input data matrix $\mathbf{x}$ is a horizontal concatenation of a matrix $\mathbf{s}$, containing structural network information, and a matrix $\mathbf{p}$, containing profile features information: $\mathbf{x} = [\mathbf{s} \mid \mathbf{p}]$. Rows represent users of the egonet and therefore for every user u there is a row vector of structural network information, $\mathbf{s}_u$, and a row vector of profile features information, $\mathbf{p}_u$. Therefore, the number of rows of the matrix $\mathbf{x}$ is the number of users in the ego-network $\mid u \mid$, and the number of columns of the matrix $\mathbf{x}$ is the total number of features used to represent structural and profile information of each user.

3.2 Structural network representation

In this subsection, we present the different representations of the structural network information that have been considered. All of them transform graph links into the matrices $\mathbf{s}$. We use the following concepts:

- *Friendship ranks:* when there is a link between two users, we say they are direct friends or rank 1 friends. When two users are not direct friends but have a common direct friend, we say they are rank 2 friends. Friendship ranks of greater levels can be further defined. In this study we consider up to rank 3 friends. There is a column in $\mathbf{s}$ for every friendship rank and user in the egonet. An element of $\mathbf{s}$ is 1 if the row user and the column user are friends of such rank, and 0 otherwise. Obtaining in total $3\times \mid u \mid$ structural features for each user.
- *Weighting:* the data is weighted depending on the friendship rank it represents. Rank 1 friendship is left with 1, whereas rank 2 friendship is weighted to 0.5 and rank 3 friendship is weighted to 0.25. Like in the previous case, obtaining in total $3\times \mid u \mid$ structural features for each user.
- *Aggregation:* for every user, the different friendship ranks are aggregated into just one value. This is obtained by calculating the maximum weighted friendship rank. Reducing the number of structural features to $\mid u \mid$.

From these concepts we define the representations shown in Table 1.

3.3 User profile representation

There are up to 57 profile features for every user in the data corpus we used for the experiments. Nevertheless, some of them are very seldom informed whereas

Table 1. Representations of structural network information

Representation	Definition
r1	Rank 1
r12	Ranks 1 and 2
r123	Ranks 1, 2 and 3
r12w	Ranks 1 and 2, weighted
r123w	Ranks 1, 2 and 3, weighted
r12a	Ranks 1 and 2, aggregated
r123a	Ranks 1, 2 and 3, aggregated

others are redundant or not relevant for the task. As a consequence, we have selected the 3 most informative features and we use only these. The selected features are: *hometown*, *schools* and *employers*. Each of these features can take different discrete values from a finite set.

We define as $\mid f \mid$ the number of features considered, and as $\mid v \mid$ the total number of values of the considered features that are taken by at least one user in the egonet. We encode the profile features information in the matrices $\mathbf{p}$, for which the following representations have been defined:

- *Explicit:* There is a column of $\mathbf{p}$ for every different value of the considered features. An element of $\mathbf{p}$ is 1 if the row user takes the column value for the respective feature, and 0 otherwise. Obtaining in total $\mid v \mid$ profile features for each user.
- *Intersection:* There is one column of $\mathbf{p}$ for every user in the egonet and every considered profile feature. An element of $\mathbf{p}$ is 1 if the sets of values of the row user and the column user, for that particular feature, intersect. It is 0 otherwise. In this case, obtaining $\mid f \mid \times \mid u \mid$ profile features for each user.
- *Weighted:* There is just one column of $\mathbf{p}$ for every user in the egonet. An element of $\mathbf{p}$ represents the proportion of features for which the row user and the column user share at least one value. It is calculated as $\frac{\mid s \mid}{\mid f \mid}$, where $\mid s \mid$ is the number of features shared between both users. Reducing the number of profile features to $\mid u \mid$.

4 Experiments

The corpus we use for the experiments is the one published for the Kaggle competition on learning social circles in networks [5]. The data consist of hand-labelled friendship egonets from Facebook and a set of 57 profile features for every node in those networks. We discarded every egonet for which the ground truth is not available. Out of the 60 egonets we finally considered, the smallest one contains 45 users and the largest one contains 670 users. The 60 egonets altogether comprise 14,519 users.

The degree of a given user is defined as the number of different circles which it belongs to. MAC takes as a parameter the range of possible degrees of the users of an egonet. In all our experiments the minimum degree is set to 0 and we try several values for the maximum degree, up to 3. In this regard, unlike previous studies, we do not include any prediction technique for the number of circles within the egonets, using the number of circles of the ground-truth instead. In future works, that would be easily incorporated with methods such as the bayesian information criterion employed in [6, 7].

The evaluation measure of our experiments, and proposed in Kaggle, is calculated as follows:
An evaluation measure for every egonet e is computed as an edit distance between the ground truth circles (g_e) and the predicted circles (p_e): $\mathrm{EDM}_e = \mathrm{d}(g_e, p_e)$. Four basic edit operations are considered: adding a user to an existing circle, creating a circle with one user, removing a user from a circle and deleting a circle with one user; every one of them at cost 1.
The evaluation measure of the whole dataset is the sum of the edit distances obtained for all the egonets.

$$\mathrm{EDM} = \sum_{e \in E} \mathrm{EDM}_e, \tag{2}$$

being E the set of the egonets in the corpus.
The smaller EDM is, the better the performance of the prediction.

5 Results

We compare our results to several different baselines. First of all, we consider MAC when it receives only structural information, using an r1 representation. MAC with only profile features, in this case, we use an explicit representation of the features. The use of both baselines has the aim to show to what degree the combination of structural network and profile information improves either of these sources of information when taken independently.

Empty circles is the third baseline we employ. This baseline relies on the fact that the evaluation measure used in this study heavily penalizes the misclassification of users into circles. Thus, defining no circle at all performs better than other possible simple baselines like connected components or classifying all the friends of an ego into just one circle.

Finally, we have considered a very high-performing baseline by using a 5-clique percolation algorithm. However, this cannot be done for every egonet due to its exponential computational complexity. Therefore, we replace the clique percolation predictions by empty circles in those cases.

It would be interesting to report results of the participants of the Kaggle competition from which we borrowed the data, as well. Unfortunately, there are only publicly available rankings for the test dataset, for which the ground truth is not available. Thus, there is no possibility to make this comparison.

The evaluation measures obtained by the baselines and our experiments are shown in Table 2. Only results obtained from weighted and aggregated structural egonet representations are presented, as non-weighting has always performed worse.

Table 2. Baselines and results of the experiments

Baseline		EDM		
MAC only structure		18679		
MAC only profile		20271		
Empty circles		17101		
Clique percolation		15350		
Data representation		EDM		
Structural	Profile	deg. 1	deg. 2	deg. 3
r12w	Explicit	16962	16803	16827
r12w	Intersection	16032	16360	15927
r12w	Weighted	17001	16955	16920
r123w	Explicit	17106	17053	17082
r123w	Intersection	16520	16504	16518
r123w	Weighted	16994	17065	17075
r12a	Explicit	15797	15619	**15570**
r12a	Intersection	16433	16840	16694
r12a	Weighted	15725	15751	15625
r123a	Explicit	16804	16770	16703
r123a	Intersection	16960	17000	17383
r123a	Weighted	16634	16542	16558

The best results have been produced when considering friendship of ranks 1 and 2, aggregated; and an explicit representation of the profile features information, allowing MAC for a maximum degree of 3. This representation has provided a value of EDM close to that obtained from the clique percolation baseline. All the experiments using the structural network representation r12a have given low values of EDM, outperforming the empty circles baseline in all the cases and most of the other representations as well. The combination of (weighted) structural network information and profile features has always performed better than structure or profile separately.

6 Conclusions

Network structure and profile features are complementary sources of information for social circles detection. In addition, weighting of structural network information with respect to friendship levels is crucial to improve the results and get close to the ones provided by methods such as clique percolation. This work opens

the door to new research in the topic, being possible future experiments the use of a greater set of profile features or better retrieved ones and the adoption of other prediction techniques or even a more in-depth study of MAC.

Acknowledgement

This work was developed in the framework of the W911NF-14-1-0254 research project Social Copying Community Detection (SOCOCODE), funded by the US Army Research Office (ARO).

References

1. Buhmann, J., Kühnel, H.: Vector quantization with complexity costs. IEEE Transactions on Information Theory 39 (4), 1133–1145 (1993)
2. Dey, K., Bandyopadhyay, S.: An empirical investigation of like-mindedness of topically related social communities on microblogging platforms. International Conference on Natural Languages (2013)
3. Fortunato, S.: Community detection in graphs. Phys. Rep. 486 (3), 75–174 (2010)
4. Frank, M., Streich, A. P., Basin, D., Buhmann, J. M.: Multi-assignment clustering for boolean data. J. Mach. Learn. Res. 13 (1), 459–489 (2012)
5. Kaggle: Learning social circles in networks. http://www.kaggle.com/c/learning-social-circles
6. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. Advances in Neural Information Processing Systems 25, 539–547 (2012)
7. McAuley, J., Leskovec, J.: Discovering social circles in ego networks. ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (1), 4 (2014)
8. Palla, G., Dernyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–818 (2005)
9. Pathak, N., DeLong, C., Banerjee, A., Erickson, K.: Social topic models for community extraction. The 2nd SNA-KDD Workshop (2008)
10. Porter, M. A., Onnela, J. P., Mucha, P. J.: Communities in networks. Notices Amer. Math. Soc. 56 (9), 1082–1097 (2009)
11. Rose, K., Gurewitz, E., Fox, G. C.: Vector quantization by deterministic annealing. IEEE Transactions on Information Theory 38 (4), 1249–1257 (1992)
12. Sachan, M., Contractor, D., Faruquie, T. A., Subramaniam, L. V.: Using content and interactions for discovering communities in social networks. Proceedings of the 21st international conference on World Wide Web, 331–340 (2012)
13. Streich, A. P., Frank, M., Basin, D., Buhmann, J. M.: Multi-assignment clustering for Boolean data. Proceedings of the 26th Annual International Conference on Machine Learning, 969–976 (2009)
14. Vaidya, J., Atluri, V., Guo, Q.: The role mining problem: finding a minimal descriptive set of roles. Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, 175–184 (2007)
15. Zhou, D., Councill, I., Zha, H., Giles, C. L.: Discovering temporal communities from social network documents. Seventh IEEE International Conference on Data Mining, 745–750 (2007)